Performance and prediction: Bayesian modelling of fallible choice in chess

Article

Accepted Version

Performance and Prediction: Bayesian Modelling of Fallible Choice in Chess

Guy Haworth[1], Ken Regan[2] and Giuseppe Di Fatta[1],

Abstract. Evaluating agents in decision-making applications requires assessing their skill and predicting their behaviour. Both are well developed in Poker-like situations, but less so in more complex game and model domains. This paper addresses both tasks by using Bayesian inference in a benchmark space of reference agents. The concepts are explained and demonstrated using the game of chess but the model applies generically to any domain with quantifiable options and fallible choice. Demonstration applications address questions frequently asked by the chess community regarding the stability of the rating scale, the comparison of players of different eras and/or leagues, and controversial incidents possibly involving fraud. The last include alleged under-performance, fabrication of tournament results, and clandestine use of computer advice during competition. Beyond the model world of games, the aim is to improve fallible human performance in complex, high-value tasks.

Keywords: assessment, behaviour, Bayes, choice, fallible, skill

1 Introduction

In the evolving world today, decision-making is becoming ever more difficult. Professionals are increasingly working as parts of man-machine systems, helped or supplanted by intelligent, carbon agents. Those responsible for the quality of the decisions therefore have a need to (a) assess the quality of their agents, and (b) predict the probabilities of other agents' choices in 'zero sum' situations. These needs are clear in real-time financial scenarios – city markets, auctions, casinos - and for effective control of utility services.

A method is proposed here for modelling and analysing decision-making in complex but quantifiable domains. The 'model world' of chess serves, as it has often done in the past, as a demonstration domain.

Skill in the global chess community has been measured by the FIDE Elo system [1] on the basis of past results. However, a good player needs to assess their opponents' skill of the moment, and chooses a move which is worst for the opponent rather than best in an absolute chessic sense. The human factor is perhaps more evident in a game of Poker or Roshambo.[3] Skill assessment is an analysis of the past, but performance prediction more dynamically considers the parameters of the current situation. One might consider that the better choices are more likely than worse ones, but that the less the apparent skill or rationality of the decision-maker, the more likely the worse choices are to be made.

[1] School of Systems Eng., Univ. of Reading, RG6 6AH, UK: g.haworth@rdg.ac.uk.
[2] Dept. of CS and Eng., Univ. at Buffalo, State Univ. of New York, Buffalo, NY 14260-2000.
[3] Roshambo is also known as *Rock, Paper, Scissors*, a pure exercise in opponent assessment.

The proposed modelling method uses a *Benchmark Space* and a *Bayesian Inference* mapping of behaviour into that space. The *space* is seeded by Reference Agents which have or are given defined dimensions of fallibility. Bayes' method is used to profile the decision-maker in terms of fallible agents. Thus, the decision-maker or agent analysed is not only positioned relative to other agents and possible threshold performance levels but is also rated in an absolute sense. The demonstration applications in the chess domain address frequently asked questions and some topical issues concerning various forms of cheating. One of these, somewhat ironically, is the illicit use of computer-advice during competition.

The Bayesian approach was first proposed [2] in the subdomain of chess where perfect information about the quality of the moves is known. Extending it to chess generally [3, 4, 5] requires resort to fallible benchmarks yielding confident rather than certain results. Nevertheless, [5] shows strong correlation between the current FIDE Elo rating scale and the new *apparent competence* rating *c*.

Section 2 defines the two concepts of Agent Space and Bayesian Inference Mapping, and notes a missed opponent-modelling opportunity. Section 3 extends the principle to that part of chess where engines evaluate positions heuristically. Section 4 reviews the application of the theory in the laboratory and to chess questions of interest. In summarizing, we anticipate the evolution of the approach, its further application in chess, and its use in non-game 'real world' scenarios.

2 Absolute Skill in the Chess Endgame

The Chess Endgame is defined here as that part of chess for which Endgame Tables (EGTs) have been computed. An EGT gives the theoretical value and Depth to Goal of every legal position for an endgame force, e.g. King & Queen v King & Rook (KQKR). The most compact and prevalent EGTs are those of Nalimov [6], providing Depth to Mate (DTM) where mate is the end-goal of chess: these are used by many chess engines on a simple look-up basis. EGTs for all required 3-, 4-, 5- and 6-man endgames are available up to KPPKPP.

Given this database of perfect information, some questions suggest themselves: a) how difficult are various endgames, b) how long might a hypothetical fallible agent take to win a game, and c) how well do humans play endgames?

Although Jansen [7] had addressed the topic of Opponent Fallibility, it was left to Haworth [2] to define an agent space SRFEP of Reference Fallible Endgame Players (RFEPs) as defined in the next section.

2.1 The Agent Space SRFEP of Reference Fallible Endgame Players

Let E be an engine playing an endgame using an EGT: further, let E have a theoretical win of depth d_0. Let $E(c)$ be a stochastic variant of E with apparent competence c, constrained to retain the win[4] but choosing its moves by the following algorithm:

[4] The extension of the model to drawing/losing moves has been done but is not needed here.

- let $\{m_j\}$ be the available winning moves, respectively, to depths $\{d_j\}$,
- move-indexing: $i < j \Rightarrow d_i \leq d_j$, i.e. lower-indexed moves are 'no worse',
- let Prob[$E(c)$ chooses move m_j] $\propto$ Likelihood[m_j] $\equiv L(j, c) \equiv (1 + d_j)^{-c}$

The space SRFEP of RFEPs satisfies the following requirements:

a. centred: $E(0)$ is a zero-skill agent – all moves are equally likely,
b. ordered: $c_1 < c_2 \Rightarrow$ E[$d \mid E(c_1)$ moves] $\geq$ E[$d \mid E(c_1)$ moves]
c. complete: $E(\infty)$ infallibly chooses a best move: $E(-\infty)$ is anti-infallible,
d. sensitive: if $d_{j+1} = d_j+1$, as $d_j \rightarrow \infty$, $L(j, c)/L(j+1, c) \rightarrow 1$ downwards, and
e. non-exclusive: all moves have a non-zero probability of being chosen.[5]

Three factors make SRFEP 1-dimensional, simplifying its use. Because chess engines consult the EGT directly, their specific search heuristics, search depths and evaluations are irrelevant. Nor is there a perceived need to generalize to $(\kappa + d_j)^{-c}$ with $\kappa > 0$.

Haworth [2] also modelled the endgame as a Markov Space and move-choice as a Markov Process to answer questions 'a' and 'b' above and to show where the more difficult depths of an endgame were.

2.2 Mapping a Player to the Agent Space SRFEP

Question 'c' was answered by rating a player *PL's* play on the basis of an observed set of moves $M \equiv \{M_i\}$. This was done by mapping *PL*, given M, to a profile of engines $\{E(c)\}$ in SFREP.

Let us suppose that the moves M_i, in fact played by *PL* in the endgame on whatever basis, have in fact been played by an engine $E \equiv E(c)$ where c is one of $\{c_k\}$ e.g. $c = 0$, …, 50. Let the initial probability that $E \equiv E(c_k)$ be $p_{k,0}$: for example, the 'know nothing' stance would set all $p_{k,0}$ to the same value.

Note now that, given that move M_1 is chosen from the moves m_{1j} available:

- Prob[$E \equiv E(c_k)$] = $p_{k,l-1}$ before move M_l is chosen on the l^{th} turn,
- $q_k \equiv$ Prob[$E(c_k)$ chooses move M_1] may be calculated as follows …
- $q_k \equiv (1 + d_l)^{-c} / \Sigma_j (1 + d_j)^{-c}$ where j ranges over the move-options available,
- Bayes' Rule defines the *a posteriori* probability $p_{k,1}$ that $E \equiv E(c_k)$ given M_l,
- As k varies across the range of engines E_α, $p_{k,l} \propto (p_{k,l-1} \times q_k)$,
- $p_{k,l} \equiv (p_{k,l-1} \times q_k) / \Sigma_\alpha (p_{\alpha,l-1} \times q_\alpha)$ where α ranges over the possible engines E_α,
- after observing all moves M_i, Prob[$E \equiv E(c_k)$] $\equiv p_{k,n} \equiv r_k$.

On the evidence of $M \equiv \{M_i\}$, player *PL* has been profiled in the agent space: it has been associated by a player-agent mapping *PA,* with $\{r_k E(c_k)\}$, a probability distribution of agents. In fact, engine *PA*(*PL*) may be defined as $\{r_k E(c_k)\}$, an engine which behaves like engine $E(c_k)$ on each move with probability r_k. We also have a metric for the absolute competence of *PL* in the competence rating $r_{PL} \equiv \Sigma r_k \times c_k$.

Given a fallible opponent *PL*, a player, especially if a computer engine, may model *PL*, predict their behaviour and exploit their apparent weaknesses accordingly [7, 8].

---

[5] The '1' in $(1 + d_j)$ ensures a non-zero denominator when $d_1 = 0$.

2.3 Adapting to the Opponent: a missed opportunity

In 1978, Ken Thompson armed his chess engine BELLE with a secret weapon, the KQKR EGT[6]. The KQ-side has a tough challenge [9, 10] with a budget of 50 moves to capture or mate, and 31 being needed in the worst case. Thompson wagered $100 that no-one would beat BELLE in the KQKR endgame [11-14] and only GM Walter Browne took up the two-game test. Browne failed in the first game but, rising to the competitive challenge, and partially informed by BELLE's KQKR-listings and a plan, he returned to recapture the Rook and his $100 just in time on the 50th move.

Haworth [2] gives details of the moves and progress in depth terms, and analyses them as above, as if the choices of some engine in the set $\{E(0), E(1), \ldots, E(50)\}$[7]. Had BELLE perceived Browne's *apparent competence* c_{WB}, it could have chosen correctly between DTC-optimal moves four times. In fact, it picked the right move just once, missing three opportunities to prolong its defence by the necessary one move.

3 Absolute Skill in Chess

Here are some categories of question that have been asked of human play:

a. Does 'Elo *E*' mean the same today as it did in years past?
b. How does player *PL*'s absolute skill vary over their career?
c. How does player *PL*'s skill compare with others' skill?
d. How do the games of tournament *T* compare with each other?
e. Is player *PL* demonstrating 'Fidelity to a Computer Agent' [15] …
   … in the context of *PL*'s (suspected) clandestine behaviour?

The answers are necessarily statistical and therefore their expected accuracy and the confidence that can be placed in them depends on the amount of data available[8] and its use. Game results and rating changes say little and conflate the behaviour of the two players. The many move-decisions potentially enable a better assessment of player performance in the context of consistent chess engine analysis.

The core idea in [3, 4] is to use chess engines as benchmark agents, assessing human competence on the evidence of their move-decisions and in the context of the engines' assessment of the options. The engines can rarely see a 'win in *n* moves' as in the endgame, and therefore indicate advantage and the consequent likelihood of a win, draw or loss in units of a Pawn. Note three complicating factors in comparison with the endgame-play rating challenge just discussed:

1. the engines' heuristic position evaluations vary from engine to engine,
2. for one engine, the evaluations usually vary with depth of search, and
3. deeper evaluations are more accurate but none are definitive.[9]

These specific questions indicate the range of questions now being addressed:

---

[6] In fact, computed to *Depth to Conversion* (DTC), i.e. *depth to capture and/or mate*.
[7] The *c*-bounds 0 and 50, and the Δc choice of 1 are such as not to over-influence the results.
[8] a result *ac* times more accurate is expected to require ac^2 more input data.
[9] Evaluations are merely substitutes for unattainable, perfect win/draw/loss information.

Re ‘a’: competence of 1971-1981 ‘Elo 2400 players’ v those of 1996-2006?
Re ‘b’: what is the profile of Victor Korchnoi’s skill over the years?
How do the best performances of World Chess Champions compare?
Guid and Bratko [16] address this using only apparent ‘move error’ as a metric.
Re ‘c’: how do the 1948 World Championship games compare with each other?
Re ‘d’: how did the players perform in a tournament: how do the games compare?
Re ‘e’: can we focus on and analyse *suspect play* at the time or later?

The next two sections are analogous to sections 2.1 and 2.2: they define a space of fallible agents and the way player *PL* is associated by a mapping *PA* with an agent profile *E* of engines in an agent space.

3.1 The Agent Space SRFP of Reference Fallible Players

Chess engines search to increasing depths rather than looking up EGTs, and vary in the heuristic position-evaluations they return, the agent space SRFP has in principle two dimensions which SRFEP does not:

1. (discrete) *search-depth*: evaluations at search-depths d_{min}, … , d_{max}, and
2. (discrete) *engine*: engines E_1, … , E_n may ‘seed’ the space SRFP.[10]

As benchmarks preferably demonstrate high-quality behaviour, these engines should have as high an Elo as possible in the various rating schemes for chess engines. The first computations reported here use SHREDDER 10 and TOGA II v1.3.1 to a modest search-depth of 10, although Regan [17] reports that TOGA II v.1.3.1 searching to depth 10 won a match[11] against CRAFTY 20.14 searching to depth 12. SHREDDER [18] is a multiple World Computer Chess Champion.

As better engines become available, one would expect the benchmark set of engines to change. For example, FRITZ 5.32 was state-of-the-art circa 1998 [19], but today one would prefer, e.g., RYBKA 3 and SHREDDER 11. For architectural (WINDOWS/LINUX and UCI[12]) and comparability reasons, the computations reported here continue with the original choices of SHREDDER 10 and TOGA II v1.3.1.

For the chess endgame, the non-negative destination depths were converted easily into positive likelihoods: the depths simply became positive denominators in the likelihood function *L*: the greater the depth, the less attractive that option for the winner. Here, position evaluations may be greater, equal to or less than zero: it seems natural to first convert these into positive numbers analogous to *depths* in the endgame. Again, the least attractive, i.e. smallest, evaluations should associate with the largest positive numbers. Thus, with $w = C(v) > 0$ being a conversion function, let

$j1 < j2 \Rightarrow$ move m_{j1} ‘is’ no worse than $m_{j2} \Rightarrow v_{j1} \geq v_{j2} \Rightarrow w_{j1} \equiv C(v_{j1}) \leq w_{j2} \equiv C(v_{j2})$

Note that function $C(v)$ potentially involves further parameters, each a dimension of the space SRFP. A caveat is also appropriate here. It is clear that some functions $C(v)$ have properties which are unrealistic in chess terms. For example, Haworth [4]

[10] To date, these dimensions are ‘null’ in our computations: $n = 1$ and $d_{min} = d_{max} = 10$.
[11] 12.5/20 ⇒ TOGA II v1.3.1 (depth 10) is ~90 ELO better than CRAFTY 20.14 (depth 12).
[12] UCI is the Universal Chess Interface [20].

suggested the function $C_1(v_j) \equiv 1 + |v_1| + |v_1 - v_j|$ but when coupled with the likelihood function $L(j, c) \equiv w_j^{-c}$ as in Section 2.1, the following unrealistic situation arises.

- Suppose moves $m_{1/2}$ are to positions with values $v > 0$ and $v_2 = -1$,
- $w_1 = 1 + v$ & $w_2 = 2 + 2v \Rightarrow L(1, c) = (1 + v)^{-c}$ & $L(2, c) = 2^{-c} \times L(1, c)$
- $\therefore \forall v$, Prob[Engine $E(c)$ chooses the better move m_1] $\equiv 1/(1 + 2^{-c})$,
- but in practice, the greater v, the more likely m_1 is to be chosen.

Therefore, Di Fatta et al [5] used a different $C(v)$:

- $C_2(v_j) \equiv w_j \equiv \kappa + |v_1 - v_j|$ with $\kappa > 0$, with $L(j, c) = w_j^{-c}$ for $E(c)$ as before,
- parameter κ, one more SRFP dimension, has so far been set to 0.1,
- n.b. $L(j, c)$ depends on v_1-v_j but not on v_1 or other v_i, but
- sensitivity requirement 'd' (in §2.1) suggests v_1 as a parameter of L,
- a correlation of (various Elo) players' apparent errors with v_1 is in plan.

The need to create functions $C(v)$ and $L(w)$ introduces the question of what $C(v)$ and $L(w)$ create the best agent-space SFRP, the one which most faithfully models the behaviour modelled. This question is considered further in the next section defining the association of player *PL* with a compound agent in SRFP. To summarise, SRFP is a space of agents or chess engines $E_i(d, \underline{\mathbf{c}})$ searching to depth d and 'dumbed down' by at least one parameter c.

3.2 Mapping a Player to the Agent Space SRFP

The following notation is useful for this section:

- player *PL's* moves $M \equiv \{M_i\}$ from positions $\{P_i\}$ are available for analysis,
- from position P_i, moves m_{ij} to positions P_{ij} are to be considered,[13]
- engine $E_k(d)$ evaluates position P_{ij} as having value v_{ijk} at search-depth d,[14]
- $C(v)$ maps positions values of any value to R^+: $v_1 > v_2 \Leftrightarrow w_1 < w_2$, and
- engine $E(d, \underline{\mathbf{c}})$ plays move m_{ij} with probability $\propto$ likelihood $L(w_{ij}, \underline{\mathbf{c}})$.

Let the hypothesis $H_{kd\underline{\mathbf{c}}}$ be that *PL's* moves are played by some engine $E_k(d, \underline{\mathbf{c}})$ which is in a 'candidate engine' subspace *CS* of SRFP. *Prior probabilities* $p_{kd\underline{\mathbf{c}}}$ are assigned to the $H_{kd\underline{\mathbf{c}}}$ before any moves are analysed. For example, $p_{kd\underline{\mathbf{c}}} = constant$ would represent the often adopted 'know nothing' initial stance but different profiles of *priors* may be used to see what the initial beliefs' long-term influences are.

Bayes' Rule is used to calculate what the *posterior probabilities* $p_{kd\underline{\mathbf{c}}}$ are (of $H_{kd\underline{\mathbf{c}}}$ being true) after observing one or more moves M_i. Let these posterior probabilities be $q_{kd\underline{\mathbf{c}}}$.[15] The Bayes Rule of Inference is simply stated:

$E_k(d, \underline{\mathbf{c}}) \in CS$, $Freq_{kd\underline{\mathbf{c}}} \equiv Prior\ Prob[H_{kd\underline{\mathbf{c}}}$ is true$] \times$ Prob$[M_i \mid H_{kd\underline{\mathbf{c}}}$ is true$]$,
Prob$[M_i \mid H_{kd\underline{\mathbf{c}}}$ is true$] \propto$ Likelihood$[E_k(d, \underline{\mathbf{c}})$ plays $M_i]$; $SumFreq = \Sigma_{CS}\, Freq_{kd\underline{\mathbf{c}}}$
$Posterior\ Prob[E_k(d, \underline{\mathbf{c}}) \mid M_i$ is played$] \equiv Freq_{kd\underline{\mathbf{c}}} / SumFreq$

---

[13] All legal moves are considered but engines only evaluate the best *MultiPV* moves precisely.
[14] To simplify the notation, some suffices will be suppressed on occasion as 'understood'.
[15] Bayes' contribution was a belief-modifying formula, obviating the need for heuristics.

Thus after modifying the initial $p_{kd\underline{\mathbf{c}}}$ to the final posterior probabilities $q_{kd\underline{\mathbf{c}}}$, Bayes' Rule has identified a composite agent or engine $E \equiv \langle q_{kd\underline{\mathbf{c}}}\, E_k(d, \underline{\mathbf{c}})\rangle$ which, by definition, decides at each move to play with probability $q_{kd\underline{\mathbf{c}}}$ as engine $E_k(d, \underline{\mathbf{c}})$. Thus, again, we have a mapping $PA : Player \rightarrow Agent$ associating players, carbon or silicon, with a profile of engines in the agent space SRFA.

If $s_k \equiv \Sigma_{d\underline{\mathbf{c}}}\, wd_d \times q_{kd\underline{\mathbf{c}}}$[16] and $r_{PL} \equiv \Sigma_k\, we_k \times s_k$, with some engine's perspectives perhaps more weighted than others but with $\Sigma_k\, we_k \equiv 1$, r_{PL} is an absolute rating for PL in the context of the benchmark used. It can therefore be used to compare players, carbon and silicon, of different playing leagues and different eras.

However, the competence of PL and $PA(PL)$ are not the same. Errors made by the benchmark engines when in fact PL makes the correct decision are seen by the engines as errors made by PL, so $PA(PL)$ will be somewhat less competent than PA. This complicates the otherwise trivial matter of putting humans and chess-engines on the same scale using games that have already been played[17] but Haworth [4] proposes a 'DGPS' approach, reducing error by identifying errors at reference points, to removing most of the error contributed by the inevitably fallible benchmark engines.[18]

Table 1. The apparent competence c, mean and stdev, with details of contributing data.

#	Player	$\mathbf{Elo_{min}}$	$\mathbf{Elo_{max}}$	Period	Games	Pos.	c_{min}	c_{max}	μ_c	σ_c	$\sigma_c * \mathbf{Pos}^{½}$
1	Elo_2100	2090	2110	1994-1998	217	12,751	1.04	1.10	1.0660	.00997	1.126
2	Elo_2200	2190	2210	1971-1998	569	29,611	1.11	1.15	1.1285	.00678	1.167
3	Elo_2300	2290	2310	1971-2005	568	30,070	1.14	1.18	1.1605	.00694	1.203
4	Elo_2400	2390	2410	1971-2006	603	31,077	1.21	1.25	1.2277	.00711	1.253
5	Elo_2500	2490	2510	1995-2006	636	30,168	1.25	1.29	1.2722	.00747	1.297
6	Elo_2600	2590	2610	1995-2006	615	30,084	1.27	1.33	1.2971	.00770	1.336
7	Elo_2700	2690	2710	1991-2006	225	13,796	1.29	1.35	1.3233	.01142	1.341

4 SRFA: Computations and Applications

The first 'SRFA' production computations inferred the *apparent competence c* of seven *Virtual Elo-e players*[19] [5]: the results show a correlation between c and FIDE Elos, and provide a context in which other inferred c may be assessed. Table 1 summarises the input data, the results and the standard deviation of the results which as expected is approximately inversely proportional to the square-root of the amount of input data.

The SRFA-computation programme is a continuing experiment: the next section is a description of how that experiment has been created and is being managed.

---

[16] The wd_d emphasise an engine's more accurate evaluations at deeper depths: $\Sigma_d\, wd_d \equiv 1$.

[17] 'PL & $PA(PL)$=$E(c)$ are Elo 2600' & 'Match $E/E(c) \Rightarrow E$ 400 Elo better' $\Rightarrow E$ has Elo 3000.

[18] Consider engine F, let $PA(F) = E(c)$, and let there be engine matches E-$E(c)$ and E-F. The match results will show the Elo difference between E, $E(c)$ and F.

[19] The *Virtual Elo-e Player* is a composite of many actual Elo e ($e = 2100 \pm 10$ etc) players.

4.1 The Computational Regime

The aims of the computation are to:

- acquire sound input data, and manage it assuredly, correctly and efficiently,[20]
- ensure that experimental results could be conveniently reproduced,
- exploit multiple computer platforms, separating job creation and commissioning,
- ensure that the engines adopted were of as high a quality as possible.

Some examples of chess-specific issues that needed to be managed:

- human players, with a win in hand, play safely rather than optimally:
  - Guid & Bratko [16] reasonably suggest ignoring positions outside [-2, 2],
- the robustness of statistical results from fallible benchmarks must be tested:
  - there was much criticism of [16] on these grounds, but
  - Guid et al [21] was only a partially successful response to this criticism.

Some examples of Bayesian Inference issues to be managed:

- probabilities need to be held in log-form to postpone underflow,
- setting priors must be consistent if moves/games are to be compared,
- care is required in setting/adapting the range/granularity of the hypotheses …
- otherwise, the prior probabilities will overly affect the posterior probabilities.

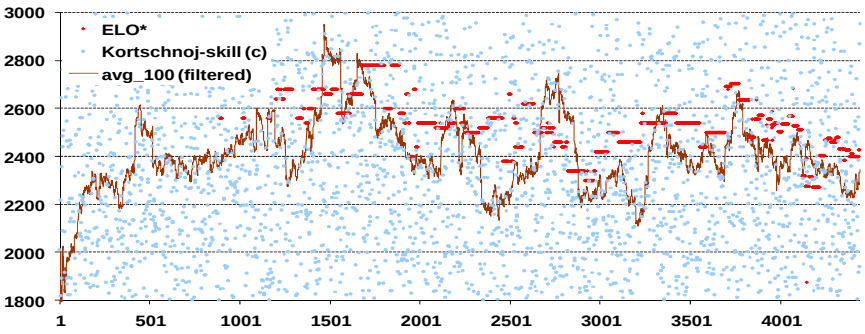

Fig. 1. Korchnoi (1950-): 'FIDE Elo bars' and *apparent competence c* over last 100 games.

4.2 Applications of 'SRFA' Computation

4.2.1 Recognised Human Achievement

Procrustes allows room here for only a sample of the insights which are now possible. Benchmarks based on reference engines enable comparison of play and players of different eras. The Elo scale is thought to have inflated [22] and a comparison of Elo 2400 play in the periods 1971-1981 and 1996-2006 is in hand. The achievements of

[20] Over 200,000 positions, their analyses and Bayesian inferences, are held in a datastore.

top players can be profiled, even before the adoption of the Elo scale in 1970: Korchnoi's *c* and Elo are shown[21] in Fig. 1. A comparison of World Champions is possible [16] though, given the quality of top-level play, the plan here is to reduce *benchmark error* and base any analysis on search-depths much greater than 10.[22]

Keres' 0-4 World Championship performance against Botvinnik in 1948 has long been a matter of speculation, as it is rumoured that he was under pressure not to impede the latter's progress to the title. Keres' and opponents' *c* per game have been computed for the 20 games in which he was involved, see Fig. 2.

Questions are asked not only about chess' finest but about the best tournaments, matches, individual performances and games on record. We look forward to identifying games where both sides played conspicuously well whatever the result.

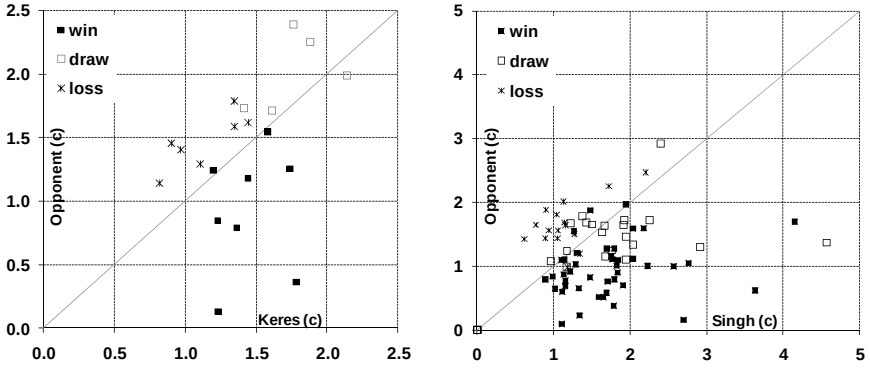

Fig. 2. Left: Keres at the WCC (1948). Right: Some 78 games by D.P.Singh (2005-8).

4.2.2 Alleged Chess Cyborgs

Players suspected of receiving computer advice during play include the following: Clemens Allwermann in 1998 [19], Diwakar Prasad Singh in 2005-6 [23, 24], Eugene Varshavsky at the World Open[23] in 2006 [25], and Krzysztof Ejsmont in 2007 [26]. In all cases, no physical evidence was found[24], the circumstantial evidence was inconclusive and probably inadequate in legal terms, and subsequent discussion of *engine similarity* lacked precision and statistical rigour. Regan [27] is addressing this lacuna and Table 2 summarises the percentage of *Move Matches* (MM) with engines' preferences for many of these scenarios. It does not yet show 'mean error' [16] but does serve as an effective sighting 'scope to target scenarios with 'SRFA'.

---

[21] The trace of Korchnoi's *c* is a running average *c* based on the last 100 games.

[22] Although fallible benchmarks give results with calculable confidence levels [4].

[23] The CCA now bans general use of mobile/(ear/head)phones and even hearing aids.

[24] Searches were instigated in the cases of Varshavsky and Ejsmont. Two players have been expelled from tournaments; Singh's colleague Umakanth Sharma was banned for 10 years.

Table 2. Frequency of player-engine Move Matches.[25]

Player	Date	Pos.	MM%	Player	Date	Pos.	MM%
Ejsmont	2007-07	104	77.6	Azmaiparashvili	1995	465	61.7
Fischer	1970+	718	67.4	Allwermann	1998-12	285	61.1
D.P.Singh	2006-04	686	64.7	SuperGMs	2005+	8447	57.5
Varshavsky/1	2006-06	170	64.2	Varshavsky/2	2006-06	44	38.3

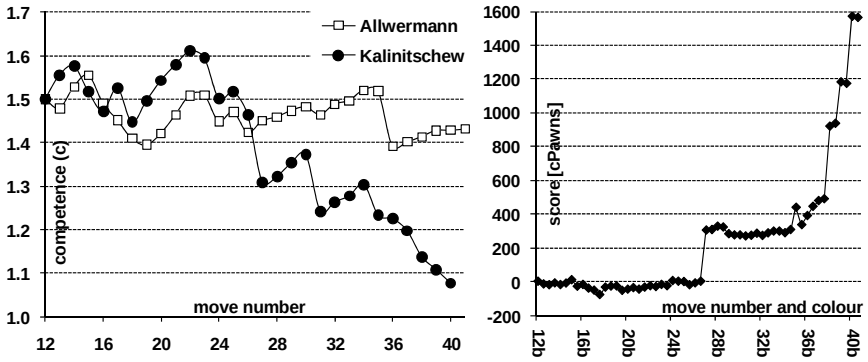

Fig. 3. Allwermann-Kalinitschew. Left: c- profile. Right: Game value in centipawns.

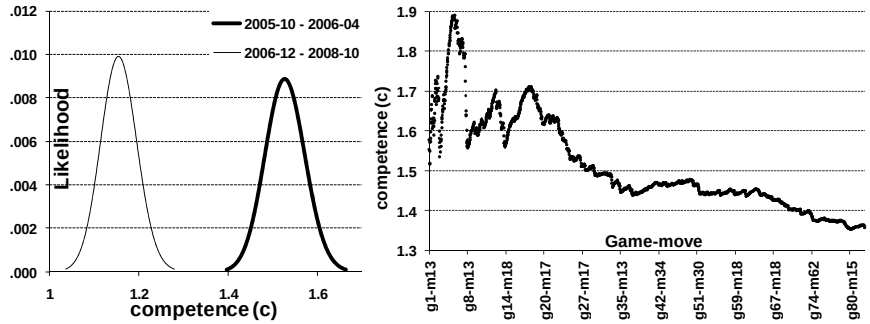

Fig. 4. D.P.Singh. Left: Probability Density Function of *apparent competence c* in two periods. Right: Evolution of *apparent competence c* based on game data available, 2005-10 to 2008-10.

[25] Varshavsky/1-2 reflects the play of this player before and after he delayed a search.

Fig. 3 addresses the performance of Allwermann and Kalinitschew in their game at the Böblingen tournament. On the left are the *c-loci* for both players, and on the right is the TOGA II v1.2.1 evaluation of the game in Pawns of advantage to White.

D.P.Singh's play came under suspicion in the second half of 2006. His *apparent competence c* profiles before and after this period are compared in Fig. 4 (left) with the evolution of his *c* alongside: the constituent games are positioned in a c_{DPS}-$c_{Opponent}$ space in Fig. 2. An application proposed here is a real-time *dashboard* (*c* plot and move series) to deter clandestine activity and to help focus the Tournament Director's forensic resources appropriately. A 'web community' implementation is feasible[26] and would also popularise chess by increasing spectator engagement and understanding.

5 The View Forward

This paper has defined and demonstrated a way of mapping decision-making behaviour into a benchmark space of agents, enabling skill to be measured in absolute terms, and future performance to be predicted.

The rating approach described here has obvious applications in identifying unexpected and possibly unwelcome behaviour. Business transactions are increasingly being carried out by/with electronic means and via the internet, facilitating the collection of evidence on the large scale necessary to reach accurate statistical conclusions. Betting markets are increasingly being monitored. The financial sector is likely to be subject to increased regulation after the collapse of trust in major institutions. The maintenance of national security increasingly seems to require the identification of patterns of electronic communication.[27]

The intention is that the Bayesian approach adopted here will be developed in several dimensions:

- Bayesian results -v- patterns of *nth-preference choice* [15],
- richer computational architecture: datastore, parallelisation, job-control,
- refined $C(v)$ & $L(w)$ functions giving better SRFP benchmark spaces,
- comparison of Bayesian results with 'average error' results [16], and
- application of the approach in one or more non-game domains.

We invite interested readers to join us in using this Bayesian approach to skill assessment, performance prediction and behaviour positioning.

Acknowledgments. The web has enabled this work to benefit from the previous initiatives of many of its contributors. We particularly thank 'SHREDDER' Stefan Meyer-Kahlen and colleague Eiko Bleicher for their continual help. We thank all those who assisted in any way, including John Beasley, Ian Bland, Paul Janota, Kerry Lawless, Soren Riis and Jeff Sonas.

[26] Trusted on-web engines send evaluations to an event server which highlights excellent play.

[27] See, e.g., the UK (RIPA) Regulation of Investigatory Powers Act (2000), the USA Patriot Act (2001) and European Community Directive 2006/24/EC on Data Retention.

References

1. Elo, A.: The Rating of Chessplayers, Past and Present. Arco (1978)
2. Haworth, G.M^{c}C.: Reference Fallible Endgame Play. ICGA J. 26-2, 81--91 (2003)
3. Haworth, G.M^{c}C.: Chess Endgame News. ICGA J., 28-4, 243 (2005)
4. Haworth, G.M^{c}C.: Gentlemen, Stop Your Engines! ICGA J. 30-3, 150--6 (2007)
5. Di Fatta, G., Haworth, G.M^{c}C., Regan, K.: Skill Rating by Bayesian Inference. In: Proc. IEEE (CIDM) Symposium on Computational Intelligence and Data Mining, 89--94 (2009)
6. Nalimov, E.V., Haworth, G.M^{c}C., Heinz, E.A.: Space-Efficient Indexing of Endgame Databases for Chess. In: Advances in Computer Games 9 (eds. H.J. van den Herik and B. Monien), pp. 93--113. IKAT, Maastricht, The Netherlands (2001)
7. Jansen, P.J.: KQKR: Awareness of a Fallible Opponent. ICCA J. 15-3, 111--131 (1992)
8. Donkers, H.H.L.M.: Nosce Hostem: Searching with Opponent Models. Ph.D. dissertation, Univ. of Maastricht (2003)
9. 'Euclid':[28] Analysis of the Chess Ending King and Queen against King and Rook. Edited by E. Freeborough. Kegan Paul, Trench, Trubner & Co. (1895)
10. Nunn, J.: Secrets of Pawnless Endings. 2nd revised edition, esp. pp. 49-69 (2002)
11. Conway, E.J.: Browne's Triumph. Chess Voice 12-2, 10 (1979)
12. Larkins, J.: Queen vs Rook: A Point for Our Side. Chess Voice 12-2, 11 (1979)
13. Stenberg, W.: Beer in the Ear. Chess Voice 12-2, 8--10 (1979)
14. Kopec, D.: Man-Machine Chess: Past, Present and Future. In: Belzer, J., Kent, A., Holzman, A.G., Williams, J.G. (eds.) Encyclopedia of Computer Science and Technology 26, 230--270, esp. 241--243. See also tinyurl.com/8faahk
15. Regan, K.W.: Measuring Fidelity to a Computer Agent. http://www.cse.buffalo.edu/-~regan/chess/fidelity/ (2007)
16. Guid, M. and Bratko, I.: Computer Analysis of World Chess Champions. ICGA J. 29-2, 65--73 (2006)
17. Regan, K.W.: http://www.cse.buffalo.edu/~regan/chess/computer/compare/Tournaments/-Toga10Crafty12Results.txt (2009)
18. Meyer-Kahlen, S.: The SHREDDER chess engine, www.shredderchess.com/ (2007)
19. Friedel, F.: Cheating in Chess. In: Advances in Computer Games 9, pp. 327-346. Institute for Knowledge and Agent Technology (IKAT), Maastricht, The Netherlands (2001)
20. Huber, R., Meyer-Kahlen, S.: The UCI Protocol. http://www.shredderchess.com/download.html (2000)
21. Guid, M., Pérez, A., Bratko, I.: How Trustworthy is CRAFTY's Analysis of Chess Champions? ICGA J. 31-3, 131--144 (2008)
22. Glickman, M.E.: Parameter estimation in large dynamic paired comparison experiments. J. Royal Stats. Soc., Series C (Applied Statistics) 48-3, 377--394 (1999)
23. Chessbase News: D.P.Singh: Supreme Talent or Flawed Genius? (2007-01-07)
24. Chessbase News: D.P.Singh has survived his 'Agni Pariksha' (2007-03-01)
25. Greengard, M.: Cheating Hearts Redux. The Daily Dirt Chess Blog (2006-07-07)
26. TWIC: Tadeusz Gniota Memorial Tournament Report. This Week in Chess (2007-07-20)
27. Regan, K.W.: Player-engine choice-matching data. http://www.cse.buffalo.edu/~regan/-chess/fidelity/FLM.html (2009)
28. Berger, J.N.: Theorie und Praxis der Endspiele, esp. p175. Revised 1922 (1890)

[28] First identified by Berger [28] as Alfred Crosskill (1829-1904) who also analysed KRBKR.